We Claim

1. A method for determining a probability for one or more states for a nucleotide in a nucleic acid sequence, comprising:

a) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in said nucleic acid sequence;

b) determining transition probabilities for each of said states for nucleotides within said nucleic acid sequence following said initial oligonucleotide;

c) determining a probability for said nucleic acid sequence for each of said states; and,

d) determining a probability for each of said states for said nucleotide based upon said probability of said nucleic acid sequence and a bias.

2. The method of claim 1, wherein said probability for each of said states for said nucleotide is determined using an inhomogeneous Markov model having eight states, wherein said eight states are: first reading frame positive strand (1+); second reading frame positive strand (2+); third reading frame positive strand (3+); first reading frame negative strand (1-); second reading frame negative strand (2-); third reading frame negative strand (3-); noncoding positive strand (N+); and, noncoding negative strand (N-).

3. The method of claim 2, wherein said probability for each of said eight states for said nucleotide in step e) is determined using the equation

$$P(f|S) = \frac{\phi(f) \cdot P_f \cdot P_f(S)}{\sum_{i \in \{1+,2+,3+,N+,1-,2-,3-,N-\}} \phi(f) \cdot P_i \cdot P_i(S)}$$

4. The method of claim 1, wherein said nucleotide is the middle nucleotide in said nucleic acid sequence.

5. The method of claim 1, wherein said nucleic acid sequence is part of a longer nucleic acid sequence.

6. The method of claim 1, wherein said bias is between 0.0 and 0.9 or greater than 1.1.

7. A method for determining a probability for one or more states for a nucleotide in a nucleic acid sequence, comprising:

a) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in said nucleic acid sequence;

b) determining transition probabilities for each of said states for nucleotides within said nucleic acid sequence following said initial oligonucleotide;

c) determining a probability for said nucleic acid sequence for each of said states; and,

d) determining a probability for each of said states for said nucleotide based upon said probability of said nucleic acid sequence, wherein said determining a probability for each of said states is capable of accepting a bias.

8. A method for determining a probability for each of one or more states for more than one nucleotide in a nucleic acid sequence comprising:

a) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in a window of a first nucleotide;

b) determining transition probabilities for each of said states for nucleotides within said window following said initial oligonucleotide;

c) determining a probability for said window for each of said states;

d) determining a probability for each of said states for said nucleotide based upon said probability for said window and a bias; and,

e) repeating steps a) through d) for each remaining nucleotide in said nucleic acid sequence.

9. The method of claim 8, wherein said more than one nucleotide are contiguous, and step e) is performed sequentially from said first nucleotide to a last nucleotide.

10. The method of claim 9, wherein said probability for each of said states for said more than one nucleotide is determined using an inhomogeneous Markov model having eight states, wherein said eight states are: first reading frame positive strand (1+); second reading frame positive strand (2+); third reading frame positive strand (3+); first reading frame negative strand (1-); second reading frame negative strand (2-); third reading frame negative strand (3-); noncoding positive strand (N+); and, noncoding negative strand (N-).

11. The method of claim 10, wherein said probability for each of said states for said more than one nucleotide is determined using the equation

$$P(f|S) = \frac{\phi(f) \cdot P_f \cdot P_f(S)}{\sum_{i \in \{1+,2+,3+,N+,1-,2-,3-,N-\}} \phi(f) \cdot P_i \cdot P_i(S)}$$

12. The method of claim 8, wherein said nucleic acid sequence is part of a longer nucleic acid sequence.

13. The method of claim 8, wherein each nucleotide in said more than one nucleotide is the middle nucleotide in its own window.

14. The method of claim 8, further comprising:

f) extending said nucleic acid sequence if said window extends beyond either end of said nucleic acid sequence, wherein said extending is accomplished by copying nucleotides from an end of said nucleic acid sequence at which said window is located to produce a copied nucleotide sequence, and adding said copied nucleotide sequence to said end.

15. The method of claim 8, wherein said window has a length of about 75 to about 125.

16. The method of claim 8, wherein said bias is between 0.0 and 0.9 or greater than 1.1.

17. A method for determining strand coding of a nucleic acid sequence, comprising:

a) determining a probability of each of one or more states for each nucleotide in said nucleic acid sequence based upon a bias, wherein each of said states is either a positive strand state or a negative strand state;

b) summing said probabilities of said positive strand states for each of said nucleotides to produce a sum of probabilities for positive states;

c) summing said probabilities of said negative strand states for each of said nucleotides to produce a sum of probabilities for negative states; and,

d) deciding one of

i) coding is mixed or not detectable if a first function of said sum of probabilities for positive states and said sum of probabilities for negative states is less than a threshold value;

ii) coding is on said positive strand if a second function of said sum of probabilities for positive states is greater than a third function of said sum of probabilities for negative states and said first function is not less than said threshold value; and

iii) coding is on said negative strand if said second function of said sum of probabilities for positive states is not greater than said third function of said sum of probabilities for negative states and said first function is not less than said threshold value.

18. The method of claim 17, wherein said sum of probabilities for positive states is X, said sum of probabilities for negative states is Y and said first function is $f(X,Y) = \frac{|X-Y|}{(X+Y)}$.

19. The method of claim 18, wherein said threshold value is from about 0.4 to about 0.6.

20. The method of Claim 17, wherein said sum of probabilities for positive states is X, said sum of probabilities for negative states is Y, said second function is f(X)=X, and said third function is f(Y)=Y.

21. The method of claim 17, wherein step a) comprises:

e) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in a window of a first nucleotide;

f) determining transition probabilities for each of said states for nucleotides within said window following said initial oligonucleotide;

g) determining a probability for said window for each of said states;

h) determining a probability for each of said states for said nucleotide based upon said probability for said window and a bias; and,

i) repeating steps e) through h) for each remaining nucleotide in said nucleic acid sequence.

22. A method for determining the extent of an open reading frame within a nucleic acid sequence, comprising:

a) determining the probability of each of one or more states for each nucleotide in said nucleic acid sequence based upon a bias, wherein each of said states is either a coding state or a noncoding state;

b) determining the coding strand of said nucleic acid sequence; and,

c) determining the points within said nucleic acid sequence in said coding strand at which the sum of the probabilities of said coding states for each nucleotide drops below a first threshold value for a number of nucleotides greater than a second threshold value, wherein ends of said open reading frame are indicated at said points.

23. The method of claim 22, wherein said first threshold value is about 0.4 to about 0.6.

24. The method of claim 22, wherein said second threshold value is about 500 to about 700.

25. The method of claim 22, wherein step c) comprises:

d) determining the sum of said coding states for a middle nucleotide located in said nucleic acid sequence;

e) repeating step d) sequentially for nucleotides located on a first side of said middle nucleotide until either

i) the sum of the probabilities of said coding states drops below said first threshold value for a number of nucleotides greater than said second threshold value, or

ii) an end of said nucleic acid sequence is reached,

at which point an end of the open reading frame is indicated; and,

f) repeating step e) for nucleotides located on a second side of said middle nucleotide.

26. The method of claim 22, wherein said nucleic acid sequence is part of a longer nucleic acid sequence.

27. The method of claim 22, wherein step b) comprises

d) summing probabilities of positive strand states for each of said nucleotides to produce a sum of probabilities for positive states;

e) summing probabilities of negative strand states for each of said nucleotides to produce a sum of probabilities for negative states; and,

f) deciding one of

i) coding is mixed or not detectable if a first function of said sum of probabilities for positive states and said sum of probabilities for negative states is less than a threshold value;

ii) coding is on said positive strand if a second function of said sum of probabilities for positive states is greater than a third function of said sum of probabilities for negative states and said first function is not less than said threshold value; and

iii) coding is on said negative strand if said second function of said sum of probabilities for positive states is not greater than said third function of said sum of probabilities for negative states and said first function is not less than said threshold value.

28. The method of claim 22, wherein step a) comprises:

d) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in a window of a first nucleotide;

e) determining transition probabilities for each of said states for nucleotides within said window following said initial oligonucleotide;

f) determining a probability for said window for each of said states;

g) determining a probability for each of said states for said nucleotide based upon said probability for said window and a bias; and,

h) repeating steps d) through g) for each remaining nucleotide in said nucleic acid sequence.

29. A method for determining the location of insertions and deletions within a nucleic acid sequence, comprising:

a) determining the probability of each of one or more states for each nucleotide in said nucleic acid sequence based upon a bias, wherein each of said states is either a coding state or a noncoding state;

b) setting a length for a window;

c) determining which state has a maximum mean probability for said nucleic acid sequence on a first side of a middle nucleotide in said window, wherein said window begins at a first nucleotide;

d) determining which state has a maximum mean probability for said nucleic acid sequence on a second side of said middle nucleotide in said window;

e) determining that a deletion or insertion occurred at said middle nucleotide if

i) said state with said maximum mean probability on said first side of said middle nucleotide is different from said state with said maximum mean probability on said second side of middle nucleotide, and

ii) either an average of hypothetical state probabilities for said window with an insertion at said middle nucleotide or an average of hypothetical state probabilities for said window with a deletion at said middle nucleotide is greater than a sum of said middle nucleotide's coding states probabilities; and,

f) repeating steps c) through e) for each remaining nucleotide in said nucleic acid sequence after said first nucleotide, wherein said window begins at each remaining nucleotide in turn.

30. The method of claim 29, further comprising:

g) determining that a deletion occurred if said average of hypothetical state probabilities for said window with an insertion at said middle nucleotide is greater than an average hypothetical state probabilities for said window with a deletion at said middle nucleotide or that an insertion occurred if said average hypothetical state probabilities for said window with an insertion at said middle nucleotide is not greater than an average of hypothetical state probabilities for said window with a deletion at said middle nucleotide.

31. The method of claim 29, wherein said nucleic acid sequence is part of a longer nucleic acid sequence.

32. The method of claim 29, wherein said repeating in step f) is performed sequentially from said first nucleotide to a last nucleotide.

sub 32 33. The method of claim 29, wherein said window is about 75 to about 125.

34. The method of claim 29, wherein step a) comprises:

g) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in a window of a first nucleotide;

h) determining transition probabilities for each of said states for nucleotides within said window following said initial oligonucleotide;

i) determining a probability for said window for each of said states;

j) determining a probability for each of said states for said nucleotide based upon said probability for said window and a bias; and,

k) repeating steps g) through j) for each remaining nucleotide in said nucleic acid sequence.

35. A method for determining exon location within a nucleic acid sequence, comprising

a) determining the probability of each of one or more states for each nucleotide in said nucleic acid sequence based upon a bias, wherein each of said states is either a coding state or noncoding state;

b) determining the coding strand of said nucleic acid sequence;

c) determining the extent of an open reading frame within said nucleic acid sequence;

d) classifying each nucleotide in a coding class or a noncoding class based on a most probable state for said coding strand;

e) reclassifying each nucleotide according to defined rules; and,

f) determining that regions of said nucleic acid sequence in said coding class are exons.

36. The method of claim 35, wherein step e) comprises:

g) reclassifying a noncoding nucleotide to a class of an adjacent nucleotide on a first side of said noncoding nucleotide and an adjacent nucleotide on a second side of said noncoding nucleotide if said adjacent nucleotide on said first side and said adjacent nucleotide on said second side all are of a single class;

h) reclassifying a nucleotide to a class of two adjacent nucleotides on a first side and two adjacent nucleotides on a second side if said two adjacent nucleotides on said first side and said two adjacent nucleotides on said second side all are of a single class;

i) reclassifying a first pair of adjacent nucleotides having a same class to a class of two adjacent nucleotides on a first side of said first pair and two adjacent nucleotides on a second side of said first pair if said two adjacent nucleotides on said first side and said two adjacent nucleotides on said second side all are of a single class;

j) reclassifying a second pair of adjacent nucleotides having a same class to a class of an adjacent nucleotide on a first side of said second pair and an adjacent nucleotide on a second side of said second pair if said adjacent nucleotide on said first side and said adjacent nucleotide on said second side both are of a single class;

k) reclassifying a nucleotide to a class of an adjacent nucleotide on a first side of said single nucleotide and an adjacent nucleotide on a second side of said nucleotide if said adjacent

nucleotide on said first side and said adjacent nucleotide on said second side both are of a single class;

l) reclassifying a continuous sequence of less than a defined minimum number of nucleotides in a noncoding class having nucleotides in a coding class on both sides to a coding class of flanking nucleotides; and,

m) reclassifying a coding segment comprising more than one class of nucleotides to a most common class in said segment.

37. The method of claim 35, wherein step b) comprises:

g) summing probabilities of positive strand states for each of said nucleotides to produce a sum of probabilities for positive states;

h) summing probabilities of negative-strand states for each of said nucleotides to produce a sum of probabilities for negative states; and,

i) deciding one of

I) coding is mixed or not detectable if a first function of said sum of probabilities for positive states and said sum of probabilities for negative states is less than a threshold value;

II) coding is on said positive strand if a second function of said sum of probabilities for positive states is greater than a third function of said sum of probabilities for negative states and said first function is not less than said threshold value; and

III) coding is on said negative strand if said second function of said sum of probabilities for positive states is not greater than said third function of said sum of probabilities for negative states and said first function is not less than said threshold value.

38. The method of claim 35, wherein step c) comprises:

g) determining the points within said nucleic acid sequence in said coding strand at which a sum of the probabilities of coding states for each nucleotide drops below a first threshold value for a number of nucleotides greater than a second threshold value, wherein ends of an open reading frame are indicated at said points.

39. The method of claim 35, wherein step a) comprises:

g) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in a window of a first nucleotide;

h) determining transition probabilities for each of said states for nucleotides within said window following said initial oligonucleotide;

i) determining a probability for said window for each of said states;

j) determining a probability for each of said states for said nucleotide based upon said probability for said window and a bias; and,

k) repeating steps g) through j) for each remaining nucleotide in said nucleic acid sequence.

40. The method of claim 35, further comprising:

g) translating said exons to determine a protein sequence.

41. A method for determining a probability for one or more states for a nucleotide in a nucleic acid sequence, comprising determining a probability for each of said states for said nucleotide based upon a probability of said nucleic acid sequence and a bias.

42. A method for determining a probability for each of one or more states for more than one nucleotide in a nucleic acid sequence comprising:

a) determining a probability for each of said states for a first nucleotide in said nucleic acid sequence based upon a probability of a window in which said first nucleotide is located and a bias; and,

b) repeating step a) for the remaining nucleotides in said nucleic acid sequence.

43. A program storage device readable by a machine, tangibly embodying a program of instructions executable by a machine to perform method steps to determine a probability for each of one or more states for a nucleotide in a nucleic acid sequence, said method steps comprising:

a) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in said nucleic acid sequence;

b) determining transition probabilities for each of said states for nucleotides within said nucleic acid sequence following said initial oligonucleotide;

c) determining a probability for said nucleic acid sequence for each of said states; and,

d) determining a probability for each of said states for said nucleotide based upon said probability of said nucleic acid sequence and a bias.

44. A program storage device readable by a machine, tangibly embodying a program of instructions executable by a machine to perform method steps to determine a probability for one or more states for more than one nucleotide in a nucleic acid sequence, said method steps comprising:

a) determining an initial oligonucleotide probability for each of said states for an initial oligonucleotide in a window of a first nucleotide;

b) determining transition probabilities for each of said states for nucleotides within said window following said initial oligonucleotide;

c) determining a probability for said window for each of said states;

d) determining a probability for each of said states for said nucleotide based upon said probability for said window and a bias; and,

e) repeating steps a) through d) for each remaining nucleotide in said nucleic acid sequence.